Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Data Mining Using K-Means Algorithm for Clustering Snack Sales at CV Sinar Pangan Utama

Bintang Tri Admaja[1]*, Arnes Sembiring[2], I Gusti Prahmana[3]

[1,3]*Informatics Engineering, STMIK KAPUTAMA Binjai, Indonesia*
[2]*Informatics Engineering, Medan Area University Medan, Indonesia*
Bintangtriadmaja321@gmail.com[1*], arnessembiring@staff.uma.ac.id[2], igustriprahmana4@gmail.com[3]

Abstract

Sales activities are a fundamental component of a company's operations in achieving profitability. CV Sinar Pangan Utama, a company specializing in the production of snacks such as morena, pang pang, amazon, and kue bawang, faces challenges related to inventory surplus and limited insights into consumer behavior. This study aims to apply data mining techniques, specifically the K-Means clustering algorithm, to analyze sales data and identify product groupings based on sales performance. By classifying products into clusters of high and low demand, the company can derive actionable insights to optimize production planning, inventory management, and marketing strategies. The research utilizes sales data spanning from January to December 2023 and is implemented using a PHP and MySQL-based application. The findings are expected to contribute to more efficient decision-making processes by uncovering purchasing patterns, thereby enhancing the company's responsiveness to market demand and improving overall business performances.

Keywords: *Clustering, K-Means, Sales Analysis, Snack Products, CV Sinar Pangan Utama*

1. Introduction

In the era of digital transformation and rapid development of information technology, data has become a highly valuable commodity for the business world. The availability of large volumes of data (big data) requires companies not only to store it, but also to analyze and manage it in order to generate relevant information as a basis for decision-making. One of the most widely used techniques to extract information from complex data is data mining. This technology enables companies to identify hidden patterns, predict trends, and optimize business operations [1]. Data mining techniques such as clustering play a crucial role in the decision-making process. The K-Means algorithm, as one of the most popular non-hierarchical clustering methods, is used to group data into several clusters based on similar characteristics [2]. The strength of K-Means lies in its simplicity and efficiency in processing large-scale data, although determining the optimal number of clusters remains a challenge [3]. The K-Means algorithm is one of the most effective and widely used methods in data mining implementation. This demonstrates that clustering techniques using K-Means offer advantages in efficiently grouping data based on similar characteristics, making it widely applied in various research and industrial fields [4]. CV Sinar Pangan Utama is a company engaged in the snack food industry, offering a variety of products such as Morena, Pang Pang, and Amazon. These products fall under the category of fast-moving consumer goods (FMCG), which are highly dependent on market dynamics and consumer preferences. Although the company already has a sales data recording system, the data has not been fully utilized to support the efficiency and effectiveness of operational and marketing strategies. One of the main problems faced is the imbalance between available stock and actual market demand, which leads to storage cost inefficiencies due to overstocking as well as missed sales opportunities caused by stock shortages. An analysis of historical sales data using a data mining approach, particularly clustering techniques with the K-Means algorithm, presents a potential solution to help the company understand product sales patterns, identify best-selling and less-preferred products, and formulate more adaptive, data-driven stock and production management strategies.

2. Literature Review

Several previous studies have demonstrated the successful application of K-Means in the context of sales analysis and product segmentation, producing 5 clusters (Cluster 0–4) with distinct characteristics [5]. A study showed the implementation of the K-Means Clustering algorithm for customer data analysis in an insurance company, which supported more personalized and efficient business strategies [6]. Another study applied the K-Means algorithm in clustering skincare products to determine marketing strategies by dividing them into 3 clusters: best-selling products, medium-selling products, and low-selling products [7].

One study used K-Means clustering to classify data based on several parameters. Cluster 0 had stock levels ranging from 500 to 900, with the peak distribution around 700, while Cluster 1 had stock levels ranging from 400 to 800, with the peak distribution around 600. In terms of price, Cluster 0 showed a peak distribution around 30,000, while Cluster 1 was around 40,000. Sales in Cluster 0 were higher and varied, approaching 1000, while Cluster 1 had lower sales, with a peak between 0 and 100. Optimization of the number of clusters based on the Elbow graph indicated the optimal number of clusters was around 4 or 5, while the Silhouette Score graph showed the highest value at k = 2, suggesting that dividing the data into two clusters was the most optimal [8].

There is also research on perfume sales clustering, where the results obtained using the K-Means algorithm showed that group C0 (best-selling products) included 9 products, group C1 (best-selling products) included 3 products, and group C2 included 13 products [9]. The clustering method using the K-Means algorithm has also been applied to population migration data by classifying each district into predefined clusters. Based on the clustering results using the K-Means algorithm, there were slight differences between manual calculation and application-based results. In Cluster 1 (C1), the manual calculation produced 7 districts, including Sumber, Bulu, Gunem, Sulang, Kaliori, Pancur, and Sluke. Meanwhile, the application produced 10 districts, including Sumber, Bulu, Gunem, Sale, Sedan, Pamotan, Sulang, Kaliori, Pancur, and Sluke. In Cluster 2 (C2), the manual calculation produced 6 districts, including Sale, Sarang, Sedan, Pamotan, Kragan, and Lasem. The application, however, produced 3 districts: Sarang, Kragan, and Lasem. Cluster 3 (C3) produced the same result in both manual and application-based calculations, with only 1 district, Rembang [10]. Another study examined talents and interests at SMK PGRI 2 Karawang, where 1105 data entries were analyzed using the K-Means Clustering Data Mining Algorithm, resulting in 26 clusters [11].

3. Research Method

This study uses the Knowledge Discovery in Databases (KDD) approach to build a model for clustering high-achieving students in Mathematics. The stages of KDD are illustrated in Figure 1 below.

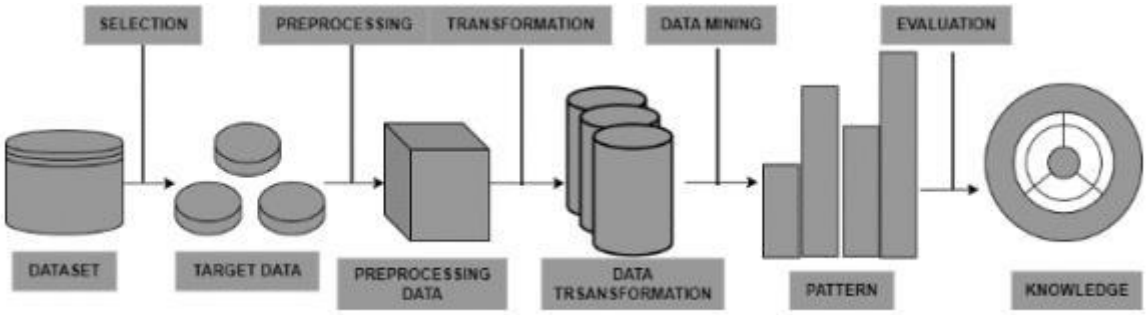

Fig. 1: Stages of KDD Method [12]

Description of Research Methods Using KDD:
1. Selection: Collecting data sales from January until December 2023 at CV Sinar Pangan Utama.
2. Preprocessing: Removing irrelevant values, addressing missing data, and normalizing grade ranges for K-Means.
3. Transformation: Converting grade data into a format suitable for the K-Means algorithm.
4. Data Mining: Using the K-Means algorithm to cluster students based on their mathematics grades.
5. Evaluation: Analyzing clustering results to assess whether the generated clusters align with the research objectives.

3.1. Data Sources

The data used in this study is sales data of snack products at CV Sinar Pangan Utama during the period from January to December 2023. The data was obtained directly from the company's administration department and has been documented in CSV format. The main attributes in this dataset include: sales date, snack brand, snack type, and sales quantity. The snack sales data can be seen in Table 1.

Table 1: Dataset

No	Sales Date	Snack Brand	Snack Type	Sales Quantity (Box)
1	2023-01-01	Amazon	Wafer	92
2	2023-01-01	Bakso Udang	Kripik	190
3	2023-01-01	Cip cip	Coklat	4
4	2023-01-01	Duo Seos	Coklat	174
5	2023-01-01	Goyang Lidah Balado	Kripik	158
6	2023-01-01	Goyang Lidah Rumput Laut	Kripik	131
7	2023-01-01	Jagung Jumbo	Kripik	79
8	2023-01-01	Kacang Arcis	Kacang	31
9	2023-01-01	Kacang Atom	Kacang	27
10	2023-01-01	Kacang Bali	Kacang	87
11	2023-01-01	Kacang Bogor	Kacang	104
12	2023-01-01	Kacang King Pedas	Kacang	24
13	2023-01-01	Kacang Mete	Kacang	90
……	……	……	……	……
……	……	……	……	……
9854	2023-12-31	Roti Kacang	Roti	172
9855	2023-12-31	Wataro Rumput Laut	Kripik	190

3.2. Data Collection Method

The data collection method in this study was carried out through:
a. Structured interviews with management to understand the data structure and sales policies
b. Documentation study, namely collecting data from the company's sales archives
c. Direct observation of the recording and distribution processes in order to validate the data

3.3. Data Analysis Technique

Data analysis was conducted through several systematic stages referring to the data mining process flow such as:

3.3.1. Data Preprocessing

This stage includes:

1. Preparing the research data
2. Data selection, which involves selecting relevant attributes such as snack type, snack brand, and sales quantity
3. Data cleaning, which involves removing duplicate or empty records
4. Data transformation. This process was carried out using a scoring system based on the regulations of CV Sinar Pangan Utama. The scoring used as the basis for transforming the data in this study is as follows [Table 2-4]

Table 2: Snack Type Data Transformation

Snack Type	Score
Amazon	1
Bakso Udang	2
Cip cip	3
Duo Seos	4
Goyang Lidah Balado	5
Goyang Lidah Rumput Laut	6
Jagung Jumbo	7
Kacang Arcis	8
Kacang Atom	9
Kacang Bali	10
Kacang Bogor	11
Kacang King Pedas	12
Kacang Mete	13
Kacang Oven	14
Kue Bawang	15
Kuping Gajah	16
Makaroni	17
Mie Goreng	18
Mie Kremes	19
Morena	20
Pang Pang	21
Potaking Balado	22
Potaking Original	23
Rasane balado	24
Rasane Rumput Laut	25
Roti Kacang	26
Watano Rumput Laut	27

Table 3: Snack Brand Data Transformation

Snack Brand	Score
Kripik	1
Roti	2
Wafer	3
Kacang	4
Coklat	5

Table 4: Sales Quantity Data Transformation

Sales Quantity	Score
1-13	1
14-26	2
27-39	3
40-52	4
53-65	5
66-78	6
79-91	7
92-104	8
105-117	9
118-130	10
131-143	11
144-156	12
157-169	13
170-183	14
184-196	15

After all the research data was changed based on the transformation table above, the following data transformation results were obtained in Table 5.

Table 5: Transformed data

Id	Snack Brand	Snack Type	Sales Quantity
1	1	2	8
2	2	1	15
3	3	5	1
4	4	5	13
5	5	1	13
6	6	1	11
7	7	1	7
8	8	4	3
9	9	4	3
10	10	4	7
11	11	4	8
12	12	4	2
13	13	4	7
……	……	……	……
……	……	……	……
9854	26	2	14
9855	27	1	15

3.3.2. Implementation of the K-Means Algorithm

Data clustering using the K-means algorithm is carried out using the steps shown in Fig 2 below.

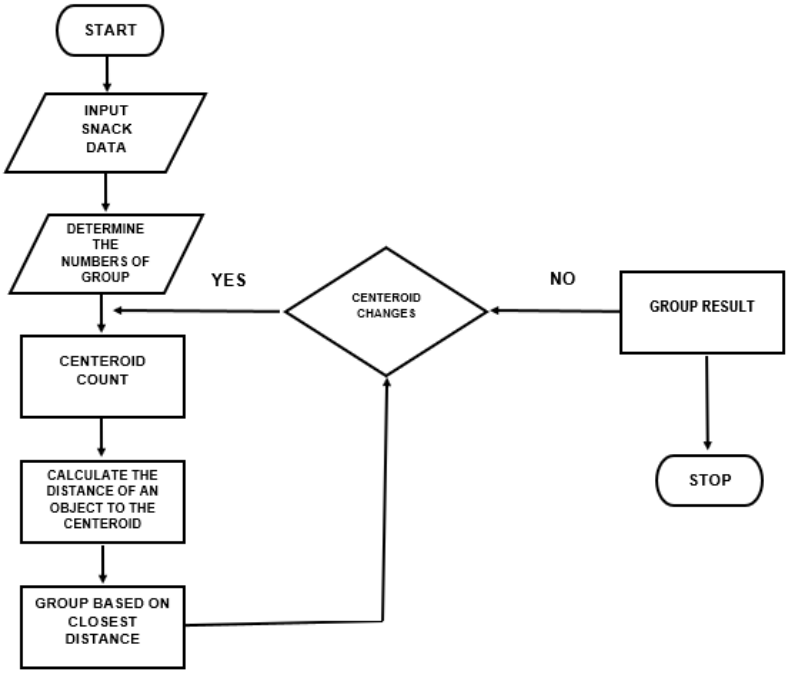

Fig. 2: K-Means Algorithm Process Flow Diagram

From the fig. 2, the flowchart process of data mining for snack product clustering using the clustering method can be described as follows:

1. Start
2. Input snack data
3. Determine the number of data groups to be processed
4. Calculate the centroid using the clustering method
5. Calculate how frequently objects are assigned to the centroid
6. Group based on snack brand, snack type, and sales quantity. Check whether the centroids are identical or not. If not, recalculate the centroid; if yes, proceed to the grouping result.
7. End process

To determine the group of an object, the first step is to measure the Euclidean distance between two object points, which is defined as follows:

$$d = \sqrt{\sum_N (Xi - Yi)^2} \quad (1)$$

Where :

d = distance (2)
X_i = Attribute data point x (3)
Y_i= Attribute data point y (4)
i = Index attribute in data (5)
N = Result of addition $(X_i - Y_i)^2$ (6)

Here is the Centroid calculation

$$\text{Centeroid} = \frac{\text{The sum of all attribute values}}{\text{amount of data}} \quad (7)$$

A centroid is the average point of a dataset, calculated by summing all attribute values and dividing by the number of data points.

4. Result and Discussion

4.4.1 Data Processing Result

After the process of data collection and cleaning, the analysis was carried out using the clustering method to group snack products based on three main variables: snack brand, snack type, and daily sales volume. The algorithm used in this study is K-Means, due to its effectiveness in grouping data into several clusters based on similarity of characteristics. In this research, sales data of snack products from CV. Sinar Pangan Utama was grouped using the K-Means algorithm with a total of five clusters.

The selection of five clusters was intended to produce a more specific and in-depth segmentation of product sales performance.

1. Cluster 1: Very High Sales Products
   Products in this cluster have very high daily sales volumes. They usually consist of leading brands that are widely recognized and have strong customer loyalty.
   Characteristics:
   a) Sales > 100 boxes/day
   b) Common snack types such as Kue Bawang or Kacang Atom
   c) Stable demand levels
   d) Suitable as a priority for distribution expansion and larger stock procurement
2. Cluster 2: High Sales Products
   Products in this cluster record high sales but not as strong as those in Cluster 1. They generally still have a good market share, though not as popular as the top-selling products.
   Characteristics:
   a) Sales 70–100 boxes/day
   b) Influenced by promotions or seasonal factors
   c) Suitable for customer loyalty improvement strategies
3. Cluster 3: Moderate Sales Products
   Products in this cluster show moderate sales. These products have potential to grow through promotional efforts or repackaging.
   Characteristics:
   a) Sales 40–69 boxes/day
   b) Growth potential exists
   c) Requires evaluation of marketing strategies
4. Cluster 4: Low Sales Products
   Products in this cluster have low sales and require more attention in terms of promotion, packaging, or pricing strategy.
   Characteristics:
   a) Sales 20–39 boxes/day
   b) Possibly less appealing to the local market
   c) Require special promotional strategies or adjustments

5. Cluster 5: Very Low Sales Products
   These products record very low sales. They are likely mismatched with consumer preferences or have uncompetitive pricing.
   Characteristics:
   a) Sales < 20 boxes/day
   b) Require comprehensive evaluation (product, pricing, distribution)
   c) Consideration for discontinuation of production/distribution

The grouping of data into five clusters provides greater granularity in analysis. Compared to a simple three-cluster division (high–moderate–low), this segmentation offers more detailed insights into each product, enabling more targeted product management strategies. The result show in Fig 3 below.

Cluster	Centeroid (x, y, z)	Jumlah Data
Cluster 1	4, 5, 12	1322
Cluster 2	6, 1, 5	1970
Cluster 3	15, 1, 12	1812
Cluster 4	16, 1, 4	1847
Cluster 5	24, 1, 8	2539

Fig. 3: Clustering Result

From the sales data from January to December 2023, the grouping of food/snack data resulted in 5 clusters: cluster 1 contains 1,322 data, cluster 2 contains 1,970 data, cluster 3 contains 1,812 data, cluster 4 contains 1,847 data, and cluster 5 contains 2,539 data.

1) Cluster 1 with total 1.322 data
   It can be seen that cluster 1 is centered on 4, 5, 12, which refers to the snack Duo Soes with daily sales ranging from 144–156 boxes.
2) Cluster 2 with total 1.970 data
   It can be seen that cluster 2 is centered on 6, 1, 5, which refers to the snack Goyang Lidah Rumput Lautdaily with sales ranging from 53–65 boxes.
3) Cluster 3 with total 1.812 data
   It can be seen that cluster 3 is centered on 15, 1, 12, which refers to the snack Kue Bawang with daily sales ranging from 144–156 boxes.
4) Cluster 4 with total 1.847 data
   It can be seen that cluster 4 is centered on 16, 1, 4, which refers to the snack Kuping Gajah with daily sales ranging from 40–52 boxes.
5) Cluster 5 with total 2.539 data
   It can be seen that cluster 5 is centered on 24, 1, 8, which refers to the snack Rasane Balado with daily sales ranging from 92–104 boxes.

4.4.2 Visualization of Clustering Results

The clustering process was further represented through graphical visualization. This visualization illustrates the distribution of data within each cluster based on sales volume and product type. The results indicate that products with similar characteristics are consistently grouped within the same cluster, thereby demonstrating the effectiveness of the clustering method in distinguishing product categories according to their sales performance. The visualization can be seen in fig 4 below.

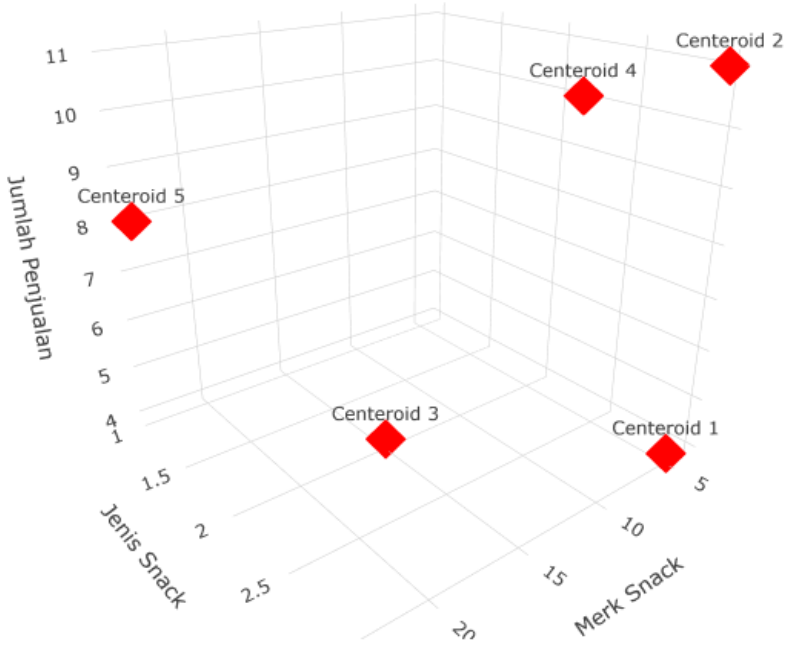

Fig. 4: Visualization of Clustering Results

4.4.3 Number of Iterations

The number of iterations can also be seen in the fig 5 below.

Iterasi	Centeroid	Jumlah Anggota per Cluster
1	(3,3,12) (5,3,6) (14,3,12) (16,2,4) (24,1,8)	577, 2323, 1921, 2380, 2289
2	(4,3,12) (6,3,5) (14,3,12) (16,2,4) (24,1,8)	1070, 1920, 1871, 1971, 2658
3	(4,3,12) (6,3,5) (15,3,12) (16,2,4) (24,1,8)	1322, 1944, 1719, 1847, 2658
4	(4,3,12) (6,3,5) (15,3,12) (16,2,4) (24,1,8)	1322, 1970, 1812, 1847, 2539

Fig. 5: Number of Iterations

4.4.4 Implications of the Five-Cluster Segmentation

The grouping of data into five clusters provides a higher level of granularity in the analysis. Compared to a simple three-cluster division (high–moderate–low), this segmentation offers a more detailed understanding of each product, thereby enabling more targeted product management strategies. The cluster characteristic can see in table 6.

Table 6: Product Management Strategies

Cluster Number	Characteristics	Recommended Strategy
1	Very High Sales	Expansion of distribution, maintain high stock levels, ensure consistent quality
2	High Sales	Continuous promotion, strengthen customer loyalty
3	Moderate Sales	Reevaluate market positioning, support with promotions and discounts
4	Low Sales	Price repositioning, repackaging, enhanced promotional efforts
5	Very Low Sales	Comprehensive evaluation, potential discontinuation of product

5. Conclusion

This study successfully applied the K-Means clustering algorithm to the sales data of snack products at CV. Sinar Pangan Utama, producing five clusters with distinct characteristics. The segmentation not only enhanced the granularity of analysis but also provided valuable insights for product management, including stock optimization, promotion strategies, product repositioning, and distribution evaluation. Compared to traditional three-cluster models, the five-cluster segmentation yielded deeper insights, enabling more strategic decision-making and improving operational efficiency. Thus, K-Means clustering proves to be an effective data mining tool for supporting data-driven business strategies in the FMCG sector.

Acknowledgement

The author would like to express gratitude to Allah SWT for His blessings and guidance that facilitated the completion of this journal. The author also extends sincere appreciation to the supervising lecturer for the invaluable assistance and guidance provided, as well as to the parents, friends, and colleagues at STMIK KAPUTAMA Binjai for their support throughout the course of this research.

References

[1] D. M. Manihuruk, H. Sabilillah dan T. Sutabri, “Big Data Analytics untuk Meningkatkan Pengambilan Keputusan di Industri”, Jurnal Pendidikan Tambusai, vol. 9, no. 1, pp. 3223-3227, 2025

[2] S. I. Murpratiwi, I. G. Agung Indrawan, and A. Aranta, “Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail,” Jurnal Pendidikan Teknologi dan Kejuruan, vol. 18, no. 2, pp.152-163, 2021, doi: 10.23887/jptk-undiksha.v18i2.37426.

[3] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and M. Jajuli, “Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM,” Jurnal Teknologi dan Informasi (JATI), vol. 12, no. 1, pp. 64–77, 2022, doi: 10.34010/jati.v12i1.6674.

[4] F. Khalish, N. M. Piranti dan O. Martadireja, “Implementasi Data Mining Menggunakan Teknik Clustering dengan Metode K-Means”, Jurnal Ilmiah Ilmu Pendidikan, vol. 8, no. 5, pp. 5392-5397, 202.

[5] H. Safitri, S. P. Lenggo Geni, F. Merry, M. Wati dan Haviluddin, “Penerapan K-Means Clustering untuk Segmentasi Konsumen E-Commerce Berdasarkan Pola Pembelian”, Jurnal Komputer dan Informatika (JUKI), vol. 7, no. 1, pp. 89-99, 2025

[6] A. A. Alya Putri and S. A. Rahmah, “Implementasi Data Mining Dengan Algoritma K-Means Clustering Untuk Analisis Bisnis Pada Perusahaan

Asuransi," Djtechno Jurnal Teknologi Informasi, vol. 5, no. 1, pp. 139–152, 2024, doi: 10.46576/djtechno.v5i1.4537.
[7] M. A Barata, I. S. Ayuni, A. Y. Kartini dan Z. Alawi, "Algoritma K-Means dalam Clustering Produk Skincare untuk Menentukan Strategi Pemasaran", Jurnal Informatika Polinema (JIP), vol. 10, ed. 3, pp. 421-427, 2025.
[8] D. Presetyo, W. Lestati dan V. Atina, "Penerapan Clustering dengan K-Means Untuk Pemilihan Menu Favorite di Tetra Coffeeshop", Jurnal Teknik Informatika dan Sistem Informasi, vol. 11, no. 3, pp. 201-2018, 2024.
[9] R. Riadi and Mesran, "Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Analisa Penjualan Parfume," Jurnal of Informatics, Electrical and Electronics Engineering. vol. 2, no. 4, pp. 138–145, 2023, doi: 10.47065/jieee.v2i4.1181.
[10] N. N. Afidah dan Masrukan, "Penerapan Metodw Clustering dengan Algoritma K-Means untuk Pengelompokan Data Migrasi Penduduk Tiap Kecamatan di Kabupaten Rembang", Prosiding Seminar Nasional Matematika 6, pp.729-738, 2023
[11] A. Supriatna, W. Dharmawan, dan C. Juliane, "Algoritma K-Means Clustering Pada Pengelompokan Minat Bakat Siswa SMK PGRI 2 Karawang", Jurnal Teknik Informatika dan Sistem Informasi, vol. 10, no. 1, pp. 38-50, 2023.
[12] M. Hilman, Martanto, A R Dikananda, A. Rifai, "K-Means Algorithm for Clustering High-Achieving Students, at Madrasah Tsanawiyah Yami Waled," 2025, Journal of Artificial Inteligence and Engineering Applications, vol. 4 no. 3 pp.1538-1547, e-ISSN: 2808-4519